The II[nd] International Conference «Advanced Information Systems and Technologies, AIST 2013»
21-24 May 2013, Sumy, Ukraine

Пошук асоціативних правил з врахуванням інформативності ознак

Зайко Т. А., Олійник А. О., Субботін С. О.
Запорізький національний технічний університет, subbotin.csit@gmail.com, http://csit.narod.ru/

***Abstract** – A method of individual informativety calculation is developed. The proposed method searches the intervals of partitioning features, reduces the degree of user interaction and reduces the probability of mining association rules that uncorrectly describe investigated objects , processes and systems*

ВСТУП

Розроблення інтелектуальних систем розпізнавання образів, діагностування, прогнозування пов'язано з необхідністю виявлення нових знань про досліджувані об'єкти та процеси. У наш час ефективним засобом інтелектуального аналізу даних є асоціативні правила [1]. При розв'язанні практичних завдань, у яких досліджувані об'єкти та процеси характеризуються дійсними атрибутами, доцільним є витягання чисельних асоціативних правил [2–4] вигляду $\langle X, v(X)\rangle \to \langle Y, v(Y)\rangle$, де $v(X)$ й $v(Y)$ – значення змінних X та Y , відповідно. Як правило, ознаки, що описують досліджувані об'єкти та процеси, мають різну інформативність [1], тому з метою витягання цікавих асоціативних правил, що адекватно описують досліджувані залежності, доцільно враховувати індивідуальну значимість ознак.

Метою дослідження є розробка методу визначення інформативності ознак у базах транзакцій для виявлення числових асоціативних правил.

МЕТОД ВИЗНАЧЕННЯ ІНФОРМАТИВНОСТІ ОЗНАК

Оскільки вихідний параметр у транзакційних базах даних, як правило, не заданий, пропонується оцінювати індивідуальну значущість ознак за допомогою параметрів, що характеризують границі областей групування екземплярів (транзакцій) у просторі ознак. Отже, для визначення індивідуальної значущості ознак пропонується виконувати кластерний аналіз, у результаті якого виділяти групи (кластери) компактно розташованих транзакцій у просторі ознак $\tau_a \in I$, I – множина усіх ознак, що входять у базу D. При цьому ознаки попередньо нормуються з метою приведення значень усіх ознак до одного діапазону, що усуне вплив величини граничних значень ознаки на її індивідуальну значущість.

У результаті кластеризації виділяється $N_{кл.}$ кластерів. Для визначення значущості кожного елемента $\tau_a \in I$ будемо оцінювати його вплив для віднесення транзакції до кожного з кластерів. Очевидно, чим менше ширина діапазону зміни значень a-ї ознаки в множині транзакцій кластера K_b ($b = 1, 2, ..., N_{кл.}$), тим більше її значущість у даному кластері. Ширину діапазону будемо оцінювати як середньоквадратичне відхилення (1):

$$\sigma_{ab} = \sqrt{\sum_{g=1}^{N_{тр.b}} \left(\overline{\tau_{ab}} - \tau_{abg}\right)^2}, \qquad (1)$$

де $\overline{\tau_{ab}}$ – середнє значення a-ї ознаки в b-му кластері; τ_{abg} – g-е значення a-ї ознаки в b-му кластері; $N_{тр.ab}$ – кількість транзакцій в b-му кластері [1].

Ознаці з мінімальним значенням величини σ_{ab} будемо присвоювати максимальне значення рангу $Rg_{ab} = |I|$ в b-му кластері, наступній по зростанню значення σ_{ab} ознаці

присвоїмо ранг $Rg_{ab} = |I| - 1$ і т.д. У випадку, якщо ознаки мають однакове значення σ_{ab}, їм присвоюються однакові значення Rg_{ab}. Ознаки, що рідко зустрічаються, із середнім значенням у групі $\overline{\tau_{ab}}$, нижче мінімально припустимого ($\overline{\tau_{ab}} < \overline{\tau_{\min}}$), вважаються неінформативними в даному кластері, внаслідок чого їм присвоюється нульове значення рангу: $Rg_{ab} = 0$. Потім для кожної a-ї ознаки τ_a складаються значення рангів по всіх кластерах (2):

$$Rg_a = \sum_{b=1}^{N_{\text{êë.}}} Rg_{ab}. \qquad (2)$$

Значущість (вага) w_a ознаки τ_a може визначається в такий спосіб:

– як відношення рангу Rg_a до суми рангів усіх ознак (3):

$$w_a = \frac{Rg_a}{\sum_{A=1}^{|I|} Rg_A}; \qquad (3)$$

– як відношення рангу Rg_a до максимального значення рангів (4):

$$w_a = \frac{Rg_a}{\max\limits_{A=1,2,\ldots,|I|} Rg_A}. \qquad (4)$$

Крім запропонованого вище підходу можна використовувати підхід, що враховує границі інтервалів розбиття ознак у кластерах. У даному методі пропонується сортувати масив значень кожної ознаки τ_a по зростанню. Ліва l_{ak} та права r_{ak} границі k-го інтервалу Δ_{ak} a-ї ознаки τ_a вибираються таким чином, щоб екземпляри (транзакції) зі значенням ознаки $\tau_a \in \Delta_{ak} = [l_{ak}; r_{ak})$ відносилися до одного кластеру K_b, а екземпляри із сусідніх інтервалів – до інших кластерів $K_c \neq K_b$.

У якості міри інформативності a-ї ознаки в транзакційній базі даних D доцільно використовувати кількість інтервалів $N_{\text{êîò. }a}$, на які розбивається діапазон її значень $\Delta_a = [\tau_{a\min}; \tau_{a\max}]$: чим менше кількість таких інтервалів, тим більше інформативність ознаки.

Тому значущість ознаки τ_a будемо обчислювати за однією з формул:

– відношення мінімальної кількості інтервалів серед усіх ознак до величини $N_{\text{инт.a}}$ a-ї ознаки (5):

$$w_a = \frac{\min\limits_{A=1,2,\ldots,|I|} N_{\text{инт.A}}}{N_{\text{инт.a}}}; \qquad (5)$$

– нормоване значення величини $N_{\text{êîò. }a}$ (6):

$$w_a = \frac{\max\limits_{A=1,2,\ldots,|I|} N_{\text{инт.A}} - N_{\text{инт.a}}}{\max\limits_{A=1,2,\ldots,|I|} N_{\text{инт.A}} - \min\limits_{A=1,2,\ldots,|I|} N_{\text{инт.A}}}. \qquad (6)$$

ВИСНОВКИ

У роботі вирішена актуальна задача автоматизації процесу виявлення числових асоціативних правил. Запропоновано метод обчислення інформативності ознак у базах транзакцій, що виділяє інтервали розбиття ознак без необхідності завдання кількості інтервалів розбиття, зменшує ступінь участі користувача та вплив його суб'єктивних оцінок на результати процесу витягання асоціативних правил, що у свою чергу знижує ймовірність виявлення асоціативних правил, які некоректно описують досліджувані об'єкти, процеси та системи.

ЛІТЕРАТУРА

[1] Engelbrecht A. Computational intelligence: an introduction / A. Engelbrecht. – Sidney : John Wiley & Sons, 2007. – 597 p.

[2] Koh Y. S. Rare Association Rule Mining and Knowledge Discovery / Y. S. Koh, N. Rountree. – New York : Information Science Reference. – 2009. – 320 p.

[3] Zhang C. Association rule mining: models and algorithms / C. Zhang, S. Zhang. – Berlin : Springer-Verlag. – 2002. – 238 p.

[4] Zhao Y. Post-mining of association rules: techniques for effective knowledge extraction / Y. Zhao, C. Zhang, L. Cao. – New York : Information Science Reference. – 2009. – 372 p.